April 1, 2000

Nucleic Acids Research, 2000, Vol. 28, No. 7 e22

Octamer-primed sequencing technology: development of prim_r identification software

Gangwu Mei and Susan H. Hardin*

Department of Biology and Biochemistry, Institute of Molecular Biology, University of Houston, Houston, TX 77204-5513, USA

Received October 12, 1999; Revised January 10, 2000; Accepted January 20, 2000

ABSTRACT

Octamer sequencing technology (OST) is a primer-directed sequencing strategy in which an individual octamer primer is selected from a pre-synthesized octamer primer library and used to sequence a DNA fragment. However, selecting candidate primers from such a library is time consuming and can be a bottleneck in the sequencing process. To accelerate the sequencing process and to obtain high quality sequencing data, a computer program, electronic OST or eOST, was developed to automatically identify candidate primers from an octamer primer library. eOST integrates the base calling software PHRED to provide a quality assessment for target sequences and identifies potential primer binding sites located within a high quality target region. To increase the sequencing success rate, eOST includes a simple dynamic folding algorithm to automatically calculate the free energy and predict the secondary structure within the template in the vicinity of the octamer-binding site. Several parameters were found to be important, including base quality threshold, the window size of the template sequence segment, and the threshold ΔG value. OST, coupled with the eOST software, can be used to sequence short DNA fragments or in the finishing assembly stage of large-scale sequencing of genomic DNA.

INTRODUCTION

General approach to DNA sequencing

DNA sequencing strategies can be divided into two major classes: directed and random. In random or shotgun DNA sequencing, a large DNA fragment such as genomic DNA is physically sheared and randomly inserted into bacteriophage or plasmid cloning vectors (1). Clones are randomly selected and sequenced and the original DNA sequence is reconstructed by computer assembly. Currently, shotgun sequencing is almost exclusively used for sequencing large DNA regions (>20 kb, e.g. genomic clones) and is considered inefficient for sequencing small DNA regions. However, since a shotgun approach typically yields >90% coverage of the original DNA fragment and leaves multiple gaps at the computer assembly stage, obtaining the final few percent of desired sequences requires a disproportionate effort and usually involves a primer-directed strategy as a critical final step for gap closure to complete the sequence project.

A directed or primer walking sequencing strategy is used not only in gap closure in large-scale sequencing of genomic DNA (2), but also as a very common and efficient approach for sequencing small DNA fragments (1–3 kb), such as cDNA clones. Primer walking utilizes oligonucleotides that anneal to the DNA template at a unique location to function as a start site for chain extension by DNA polymerase. Such an approach requires a fragment of known sequence, either a standard vector sequencing primer site or a target DNA site, to design the binding primer. The sequence information obtained from the initial sequencing reaction is used to determine the primer binding site for the next sequencing reaction. These steps are repeated until the complete sequence is determined. Because such a primer-based strategy involves repeated sequencing steps from known into unknown DNA regions, the process has the advantage of minimizing the generation of redundant sequence data. Also, it does not require multiple cloning steps or subsequent template preparations.

Thus, a directed sequencing strategy provides potential advantages, i.e. higher efficiency, less redundancy, fewer cloning steps, and straightforward data assembly. However, primer walking requires that a new oligonucleotide is synthesized for each round of sequencing. The necessity of designing and synthesizing new primers, coupled with the expense and the turn-around time for synthesis of these primers, has limited the routine application of primer walking for sequencing large DNA fragments.

To eliminate the requirement for custom primer synthesis for each reaction, researchers have proposed developing a high throughput, primer-based DNA sequencing strategy that is dependent on primers selected from a pre-synthesized primer library (3–5). This primer library would reduce the cost of synthesizing specific primers for each sequencing reaction and save the time involved in designing and synthesizing these new primers.

An octamer primer library was determined to be optimal, since an octamer is the shortest primer that reliably generates cycle sequencing data and provides sufficient specificity in

*To whom correspondence should be addressed. Tel: +1 713 743 2686; Fax: +1 713 743 2636; Email: shardin@uh.edu
Present address:
Gangwu Mei, Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Room N1521, Houston, TX 77030, USA

sequencing reactions using double-stranded DNA templates as large as 20 kb (6). A complete octamer library would include 4^8 (65 536) primers, although not all octamers produce high quality sequence data. The identification of a subset of octamers, characterized by both efficient priming and frequent occurrence, and the optimization of octamer-primed reaction chemistry made possible the practical utilization of a primer library (6–8). However, after implementing the use of an octamer library, we recognized the need for a computer program to automatically identify candidate primers for subsequent sequencing reactions. Therefore, we developed the computer analysis components required for efficient library use.

This software, electronic OST (eOST), integrates the base calling software PHRED in order to identify high quality sequence regions. Integration with base calling software eliminates the time and labor cost associated with manual editing of target sequences and provides a quality assessment for each base. Additionally, since template secondary structure impacts OST reaction success (7), an analysis of the strengths of predicted structures was incorporated into this octamer identification software.

To increase the OST success rate, eOST was also used to determine several parameters that improve sequence quality, including the base quality threshold, the window size of the template sequence segment, and the threshold point of the ΔG value. Furthermore, parameters were refined to improve the correlation between the stability of secondary structure of the template around the priming site and the success or failure rate of the sequencing reaction. Thus, eOST identifies the optimal octamer primers to use in subsequent sequencing reactions. By combining speed and reagent cost savings with minimized human intervention, eOST has become an integral component of OST.

MATERIALS AND METHODS

Operating system for computation

A Macintosh operating system is coupled with an ABI PRISM 377 DNA Sequencer to collect the raw data from the sequencer. The raw data files are transferred to a UNIX operating system. The base calling software PHRED (9–10), installed on a UNIX system, is used to determine the base quality value for each base in the sequence. Based on our algorithm, a computer program, eOST, was written and compiled in the C language. The present version of eOST runs on a UNIX operating system (Sun Solaris 2.5.1).

Octamer sequencing reactions

Octamer-primed sequencing reactions were performed as described in Jones and Hardin (7).

Octamer library nomenclature

The library size was reduced from 7168 to 970 members (8). The resulting sequences of the library primers are stored in two files with three fields in each file: primer number (index in the octamer primer library), primer ID or primer name, and sequence. Each of the files contains 485 members; one file stores the primers with index 1–485, and the other contains the corresponding complementary octamer primers indexed from 486 to 970. For example, primer number m in the first file is the complement of primer number $485+m$ in the second file. If there are changes in the octamer library, the user can make the corresponding change in this file instead of rewriting the eOST code.

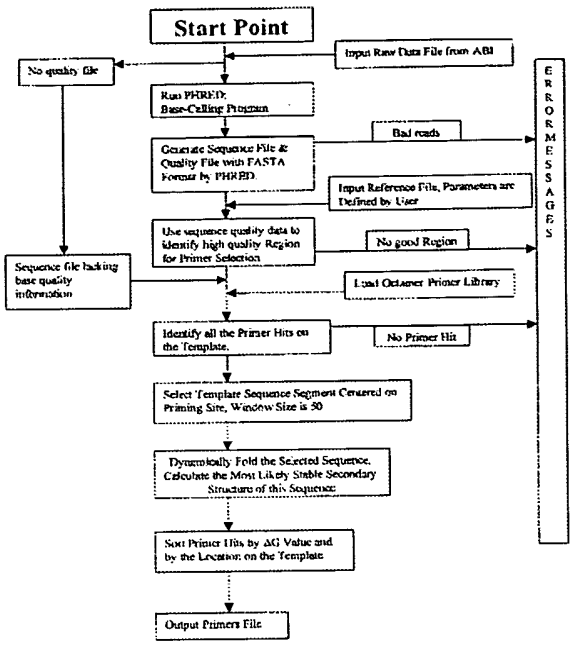

Figure 1. eOST schema.

eOST schema

An overview of the eOST design is shown (Fig. 1) and is discussed in detail in the following sections.

Using PHRED base calling software. Sequencing reaction conditions have been identified in which octamers prime reactions only from perfectly matched templates (8). A major source of reaction failure occurs when the primer is derived from an unreliable sequence (i.e. errors in the sequence are translated to the primer which fails to bind to the template), making it critical to consider the quality of the underlying sequence before selecting a primer. Since sequence quality can vary dramatically due to poor quality template, limitations of the base calling software or decay in sequence data quality near the end of the read length, it is critical to consider the quality of the underlying sequence before selecting an octamer.

To incorporate this consideration into the eOST automated octamer identification program, we have incorporated the base quality statistics generated by the PHRED base calling software to indicate the reliability of each base call (9,10). The PHRED numerical indication of base quality is used to eliminate low quality sequences that may be unreliable, and does not require viewing the trace display. Our default quality value cut-off is a PHRED value of 20, which implies that there is a 1% chance that the base call is not correct. This value can, however, be

modified by the user if more or less stringent criteria are desired.

PHRED base quality evaluation begins with the trace file collected by the automated sequencer (9,10). A set of four traces, one for each of the four bases of DNA, is stored in a binary output trace file by the Perkin-Elmer/ABI software. PHRED reads this DNA trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to the output file (9,10). After calling bases, PHRED writes the sequence and the quality values for each base to separate files in FASTA format (9,10).

Identifying a high quality template region. The two files written by PHRED, a sequence file that contains the template sequence and a quality file that contains the corresponding integer quality values for each base in the sequence, are used to identify the region for primer selection (2). Each template in the sequence file is clipped to determine a single good region for primer selection. The clipping algorithm works as follows: a sequence window of specified size (in base pairs) moves inward from the edge of the template and stops when there are less than a specified number of low quality bases in the window (2). Bases with quality below the minimum acceptable quality value (cut-off value 20) specified in the preference file are considered bad. Regions are further designated either 'good', for a high quality sequence, or 'bad', for a low quality sequence. The effect of this algorithm is to exclude runs of low quality sequence at the leading and trailing edges of individual reads (2). For each good region, the code then identifies all the candidate octamer primers from the octamer library.

Identifying candidate octamer primers. eOST analyzes the high quality sequence region for matches with octamers in the library, using alignment settings that required a 100% match and an 8 base overlap. The search stops when all candidate primers are selected and when the search region is exhausted. eOST next analyzes these candidates to determine how many times they occur in the template (within the known sequence). After a sequence is fully analyzed, the identified primers are output to a 'primers' file with the primer name, sequence, position in the template, and times present in the template. After the candidate primers are identified, template regions centered around the octamer-binding site are stored in memory for analysis of potential secondary structures.

Computing the stability of template secondary structure. There is a correlation between predicted local template secondary structure and OST reaction success or failure (7). Several factors may influence the efficiency with which a homologous DNA sequence can act as a primer for enzymatic sequencing, including poor accessibility of the primer to the template due to template secondary structure, inefficient extension by the polymerase, and multiple site binding. Thus, a better understanding of the relationship of the primary DNA sequence to the template secondary structure and its effect on primer extension by DNA polymerase should reduce the incidence of reaction failure through a more informed choice of candidate primers.

It is well established that under a given set of solution conditions, the relative stability of a DNA duplex structure depends on its base sequence (11). More specifically, the stability of a DNA duplex appears to depend primarily on the identity of the nearest neighbor bases. The overall stability and the melting behavior of any DNA duplex structure can be predicted from its primary sequence, making it possible to calculate the stability of predicted secondary structures within the template. A DNA folding algorithm, which was used to determine self-complementarity of oligonucleotides, was introduced to compute the stability of potential secondary structures (12). The specific template region surrounding the primer location was used to compute the stability of predicted secondary structures in the immediate vicinity of the candidate primer.

The template was analyzed for potential inhibitory secondary structures using an algorithm that identifies regions of self-complementarity and determines the free energy values for these regions (12). A segment of the template can theoretically be folded in many different ways. Each folding stage is associated with a free energy value. The lower the energy value, the higher the secondary structure stability of that template region. The lowest free energy for each priming site is listed in the data output file. At each priming site, a template segment of 50 bases is selected (reasons for specifying 50 bases are detailed below) and all the different folding stages are calculated.

eOST uses a simple dynamic folding strategy to generate all possible potential stem–loop structures. Folding the 5′-terminus toward the 3′-terminus and increasing the amount of overlap by a single base generates each stem–loop structure. Base pairing is analyzed as eOST progressively moves from the 5′-end towards the 3′-end. As an example, with position P of this segment of the sequence, base pairing between P and P+1 is examined. If base complementarity is observed, then P is complementary to P+1. Then, similarly, eOST continues checking the base pairing between P–1 and P+2, keeping track of the number of consecutive bases that are complementary. The base pairing process is repeated until a mismatch is identified. The length of continuous complementarity defines the length of the stem in the stem–loop structure. If the size of the stem is equal to or greater than three, then this stem–loop structure is considered significant and the free energy (ΔG) value for stability of this stem–loop structure is calculated. All the information related to this stem–loop structure is stored in data structure in memory. This information includes stem base, stem size, loop size, and location of the stem in the selected segment of template sequence. The folding process is repeated by single base shifts to generate all possible simple stem–loop structures within the entire template sequence region. Using this folding strategy, each 'hairpin structure' stored in memory is considered significant.

Both the stem and the loop impact on the stability of the hairpin structure. Information stored in memory for each structure includes the duplex length, duplex sequence, and loop size. eOST uses this information to calculate the ΔG value for each structure. The ΔG value for each stem–loop structure is considered to be the sum of the ΔG values of the duplex and the loop. First, the DNA stability of the stem in the hairpin structure is calculated using nearest neighbor thermodynamic values determined by Breslauer *et al.* (11). The second part of the calculation determines the ΔG value of the loop in the stem–loop structure (as per OLIGO 5.0; Molecular Biology Insights Inc.). Both the ΔG value and the position of each predicted stem–loop structure are stored in memory. Subsequently, these data are sorted by the ΔG values of all the possible stem–loop structures, with the

lowest ΔG value for each position being the most likely stable structure for that template region.

eOST default parameters. Based on a statistical analysis of reaction successes and failures with associated ΔG values, the best predictor of reaction outcome uses a 50 base template window (centered on the primer) with a ΔG value of –5.0 kcal/mol (addressed below in detail). The user can identify a primer from the candidate primer list whose ΔG value is in the working category. If there are two or more of these primers, the user selects the most appropriate primer by comparing the location of the priming sites in the template.

RESULTS

Correlation between stability of predicted secondary structure and OST reaction success

eOST software was tested on a sequence that was previously determined using OST (7). The template for this sequence is a double-stranded plasmid containing 4993 bp of the *Drosophila melanogaster timeless* upstream region, and is referred to as the Re2 project.

After completing the Re2 project, we found that five octamer primer pairs occurred twice in the Re2 template and that three of these five pairs were used to determine sequence information (7). Surprisingly, five of these six octamers produced high quality sequence information, indicating that they preferentially initiated the reaction from a single site, despite the fact that two primer-binding sites were present in the template. By comparing the local template ΔG values of these primers, we noted that local secondary structure differences could explain the dramatic variation in the priming efficiency of the same octamer at different priming sites in the same template (7). A correlation was seen between the site from which sequence information was primed and a low potential for hairpin formation in the bases surrounding the octamer-priming site. Subsequently, similar observations were reported by researchers performing modified, isotopic octamer-primed sequencing (13).

The observation that template secondary structure could impact on priming ability led us to closely examine the correlation between predicted secondary structure and priming extension efficiency. This correlation between predicted hairpin strength and failure of the OST reaction is supported by additional analysis of the Re2 project. A total of 63 OST reaction results are listed and sorted by the predicted lowest ΔG value (Table 1, left) or by reaction success or failure (Table 1, right). The ΔG values determined for 49 of 63 primer sites were greater than –5.0 kcal/mol (i.e. weaker secondary structure predicted) (Table 1). In this group, 42 of 49 OST reactions generated sequence data, while seven of 49 OST reactions failed. The ΔG values determined for 14 of 63 primer sites were less than –5.0 kcal/mol (i.e. stronger secondary structure predicted) (Table 1). Eleven of these 14 OST reactions failed, while three OST reactions in this group succeeded. There is a correlation between predicted free energy value and reaction success or failure (Table 1B). Also, the threshold ΔG value is around –5.0 kcal/mol. Thus, in the current version of eOST, this threshold is used as a critical point to predict reaction success or failure. The user can examine and reset the threshold ΔG value after collecting more OST reaction data.

Effect of the window size of the template segment on eOST prediction

As previously stated, eOST was applied in the Re2 project for selection of candidate primers from an optimal octamer library. Sixty-three octamer primers were selected for use in sequencing reactions without calculating free energy values of the target template surrounding the primer-binding sites. Of these octamers, 45 generated high quality sequences with a sequencing success rate of 71%, while 18 failed to generate sequence data.

To examine the effect of template window size on the free energy value, eOST was used to calculate this value at the potential primer-binding site for these same 63 primers using different template window sizes. For simplicity, a ΔG value of –5 kcal/mol was used to predict reaction success or failure (displayed in Table 2). This analysis enabled us to identify the optimal template window size associated with the greatest predictive value. Using a 35 base window size, 56 octamers with $\Delta G > -5$ kcal/mol were predicted to generate sequence information. The experimental results produced 44 (79%) successful OST reactions and 12 failed reactions. While seven octamers with $\Delta G < -5$ kcal/mol were predicted to fail, the experimental results produced six (86%) failed OST reactions and one successful reaction. Using a 50 base window size, 49 octamers with $\Delta G > -5$ kcal/mol were predicted to generate sequence data. The experimental results produced 42 (86%) successful OST reactions and seven failed reactions. Fourteen octamers with $\Delta G < -5$ kcal/mol were predicted to fail. The experimental results produced 11 (79%) failed OST reactions and three successful reactions. Using a 75 base window size, 28 octamers with $\Delta G > -5$ kcal/mol were predicted to generate sequence data. The experimental results produced 24 (86%) successful OST reactions and four failed reactions. Thirty-five octamers with $\Delta G < -5$ kcal/mol were predicted to fail, but the experimental results produced 14 (40%) failed OST reactions and 21 successful reactions. Using a 100 base window size, 24 octamers with $\Delta G > -5$ kcal/mol were predicted to generate sequence information. The experimental results produced 22 (92%) successful OST reactions and two failed reactions. Thirty-nine octamers with $\Delta G < -5$ kcal/mol were predicted to fail, but the experimental results produced 16 (41%) failed OST reactions and 13 successful reactions.

These results demonstrate that increasing the window size to select candidate primers increases the sequencing success rate and the number of primers with calculated $\Delta G < -5$ kcal/mol. However, the number of octamers that are predicted to fail but which still generate sequence data increases. To reach an optimal balance between selectivity and sensitivity, a window size of 50 was chosen as an optimal parameter for primer selection. At this window size eOST provides a high success rate of 86% (compared to 79% with window size 35), as well as a high predicted failure rate of 79% (compared to 40 and 41% for window sizes 75 and 100, respectively).

DISCUSSION

We have adapted octamer primer-based sequencing to a format that is more compatible with automation. The ultimate goal is to maximally automate the sequencing process, thereby minimizing human intervention, reducing labor costs, and accelerating the

Table 1. Re2 project summary

Primer Location	Primer Sequence	Primer Name	Number of Occurrences	eOST ΔG	OST Result
1459	GCCGATCC	730	1	0.8	1
3526	GGATCGGC	245	1	0.8	1
963	CTGTGGGG	411	1	-0.3	1
4022	CCCCACAG	896	1	-0.3	1
2355	CCAGAGGC	913	1	-0.6	1
2435	GCTTCCCC	181	1	-0.6	1
2630	GCCTCTGG	428	1	-0.6	1
2550	GGGGAAGC	666	1	-0.6	0
1926	GTGCTGGC	553	1	-0.8	1
3059	GCCAGCAC	68	1	-0.8	1
3067	GCAGCCAG	801	1	-0.8	1
369	CAGGGAGC	463	1	-0.9	1
1778	CCCACTCC	884	1	-0.9	1
3207	GGAGTGGG	399	1	-0.9	1
4616	GCTCCCTG	948	1	-0.9	1
660	CCAGGCAG	442	1	-1.1	1
1811	GTGCAGCC	567	2	-1.1	1
2207	CCCACCAG	412	1	-1.1	1
2778	CTGGTGGG	897	1	-1.1	1
3174	GGCTGCAC	82	2	-1.1	1
4325	CTGCCTGG	927	1	-1.1	1
1128	CAGTGGCC	446	2	-1.9	1
3857	GGCCACTG	931	2	-1.9	1
4696	GGCGAGAG	823	1	-2.2	1
289	CTCTCGCC	338	1	-2.2	0
299	GCTGCTCC	720	1	-2.2	0
4686	GGAGCAGC	235	1	-2.2	0
680	GAGGGCTC	270	1	-2.4	1
4305	GAGCCCTC	755	1	-2.4	1
2462	GACTCGGG	851	1	-2.5	1
2523	CCCGAGTC	366	1	-2.5	0
876	GTCCGGCA	117	2	-3.1	1
1872	GGCACAGC	232	1	-3.1	1
3113	GCTGTGCC	717	1	-3.1	1
4109	TGCCGGAC	602	2	-3.1	1
4138	CAGAGCCG	486	1	-3.1	1
847	CGGCTCTG	1	1	-3.1	0
2057	CCACAGCC	438	1	-3.3	1
2928	GGCTGTGG	923	1	-3.3	1
1894	CTGGAGGG	891	1	-3.4	1
3091	CCCTCCAG	406	1	-3.4	1
519	GTGGCCAG	88	1	-3.6	1
627	GGGCAAGG	176	1	-3.6	1
4358	CCTTGCCC	661	1	-3.6	1
4466	CTGGCCAC	573	1	-3.6	1
2410	GTTGCCCC	702	1	-3.8	1
2575	GGGGCAAC	217	1	-3.8	1
4619	GTGCGGCA	570	1	-4.8	1
366	TGCCGCAC	85	1	-4.8	0
475	GTCTGGGG	95	1	-5.1	0
4510	CCCCAGAC	580	1	-5.1	0
1310	CTGGGACC	155	1	-5.2	1
3675	GGTCCCAG	640	1	-5.2	1
2965	GGCTGCAC	82	2	-5.4	1
2020	GTGCAGCC	567	2	-5.4	0
1411	CCTGGAGC	349	1	-6.1	0
3574	GCTCCAGG	834	1	-6.1	0
324	TGCCGGAC	602	2	-6.3	0
4661	GTCCGGCA	117	2	-6.3	0
2387	CAGTGGCC	446	2	-7.2	0
2598	GGCCACTG	931	2	-7.2	0
716	GGCAGTGC	748	1	-7.3	0
4269	GCACTGCC	263	1	-7.3	0

Primer Location	Primer Sequence	Primer Name	eOST ΔG	OST Result	Previously Predicted Template ΔG *
1459	GCCGATCC	730	0.8	1	-2.7
3526	GGATCGGC	245	0.8	1	-2.7
963	CTGTGGGG	411	-0.3	1	-9.1
4022	CCCCACAG	896	-0.3	1	-9.1
2355	CCAGAGGC	913	-0.6	1	-2.6
2435	GCTTCCCC	181	-0.6	1	-3.8
2630	GCCTCTGG	428	-0.6	1	-2.6
1926	GTGCTGGC	553	-0.8	1	0.8
3059	GCCAGCAC	68	-0.8	1	0.8
3067	GCAGCCAG	801	-0.8	1	N/A
369	CAGGGAGC	463	-0.9	1	-0.9
1778	CCCACTCC	884	-0.9	1	-0.9
3207	GGAGTGGG	399	-0.9	1	-0.9
4616	GCTCCCTG	948	-0.9	1	-0.9
660	CCAGGCAG	442	-1.1	1	N/A
1811	GTGCAGCC	567	-1.1	1	-2.1
2207	CCCACCAG	412	-1.1	1	-6.0
2778	CTGGTGGG	897	-1.1	1	-6.0
3174	GGCTGCAC	82	-1.1	1	-2.1
4325	CTGCCTGG	927	-1.1	1	N/A
1128	CAGTGGCC	446	-1.9	1	-5.1
3857	GGCCACTG	931	-1.9	1	-5.1
4696	GGCGAGAG	823	-2.2	1	-4.3
680	GAGGGCTC	270	-2.4	1	-7.3
4305	GAGCCCTC	755	-2.4	1	-7.3
2462	GACTCGGG	851	-2.5	1	-2.5
876	GTCCGGCA	117	-3.1	1	-3.1
1872	GGCACAGC	232	-3.1	1	-4.1
3113	GCTGTGCC	717	-3.1	1	-4.1
4109	TGCCGGAC	602	-3.1	1	-3.1
4138	CAGAGCCG	486	-3.1	1	-3.1
2057	CCACAGCC	438	-3.3	1	-3.3
2928	GGCTGTGG	923	-3.3	1	-3.3
1894	CTGGAGGG	891	-3.4	1	N/A
3091	CCCTCCAG	406	-3.4	1	N/A
519	GTGGCCAG	88	-3.6	1	N/A
627	GGGCAAGG	176	-3.6	1	N/A
4358	CCTTGCCC	661	-3.6	1	N/A
4466	CTGGCCAC	573	-3.6	1	-3.6
2410	GTTGCCCC	702	-3.8	1	-3.8
2575	GGGGCAAC	217	-3.8	1	-3.8
4619	GTGCGGCA	570	-4.8	1	-5.7
1310	CTGGGACC	155	-5.2	1	-4.1
3675	GGTCCCAG	640	-5.2	1	-4.1
2965	GGCTGCAC	82	-5.4	1	-5.4
2550	GGGGAAGC	666	-0.6	0	-3.8
289	CTCTCGCC	338	-2.2	0	-4.3
299	GCTGCTCC	720	-2.2	0	-6.3
4686	GGAGCAGC	235	-2.2	0	-6.3
2523	CCCGAGTC	366	-2.5	0	-2.5
847	CGGCTCTG	1	-3.1	0	-3.1
366	TGCCGCAC	85	-4.8	0	-5.7
475	GTCTGGGG	95	-5.1	0	N/A
4510	CCCCAGAC	580	-5.1	0	N/A
2020	GTGCAGCC	567	-5.4	0	-5.4
1411	CCTGGAGC	349	-6.1	0	N/A
3574	GCTCCAGG	834	-6.1	0	N/A
324	TGCCGGAC	602	-6.3	0	-6.3
4661	GTCCGGCA	117	-6.3	0	-6.3
2387	CAGTGGCC	446	-7.2	0	-7.2
2598	GGCCACTG	931	-7.2	0	-7.2
716	GGCAGTGC	748	-7.3	0	N/A
4269	GCACTGCC	263	-7.3	0	N/A

Primer location, sequence, name, number of occurrences in the Re2 project, eOST predicted free energy (ΔG) value (kcal/mol), and OST experimental result are indicated. The experimental results are shown as 1 or 0, where 1 indicates OST reaction success and 0 indicates reaction failure. (Left) Sorted by eOST calculated ΔG value and then by experimental result. (Right) Sorted by experimental result and then by eOST predicted ΔG value. ΔG was calculated by eOST using a 50 base window. *, previously predicted template ΔG lists values determined using template windows sizes of 75–100 bases (7). Primer names may differ from those previously reported, however, octamer sequences can be compared. N/A indicates that the primer was not previously analyzed.

whole sequencing process. The development of eOST is another step towards this goal.

eOST was developed to automatically identify candidate primers from an on-hand octamer primer library for use in subsequent sequencing reactions. This program has been used as a tool to identify octamer primers for primer walking sequencing of a DNA template. eOST integrates the base calling software PHRED to identify candidate primer-binding sites located within a high quality target sequence. However, if the user has an accurate sequence, eOST is able to accept as input the sequence without a quality file. In fact, a version of eOST that accepts and analyzes sequence files in

Table 2. Re2 project summary of ΔG values determined using different template window sizes

Primer Location	Primer Sequence	Primer Name	Window 35 ΔG	Window 35 eOST	Window 50 ΔG	Window 50 eOST	Window 75 ΔG	Window 75 eOST	Window 100 ΔG	Window 100 eOST	OST Result
4696	GGCGAGAG	823	-0.9	1	-2.2	1	-10.4	0	-10.4	0	1
4619	GTGCGGCA	570	-4.8	1	-4.8	1	-5.7	0	-5.7	0	1
4616	GCTCCCTG	948	-0.6	1	-0.9	1	-0.9	1	-1.4	1	1
4466	CTGGCCAC	573	-3.6	1	-3.6	1	-3.6	1	-4.8	1	1
4358	CCTTCCCC	661	-2.0	1	-3.6	1	-3.6	1	-6.0	0	1
4325	CTGCCTGG	927	-1.1	1	-1.1	1	-1.7	1	-2.4	1	1
4305	GAGCCCTC	755	-1.7	1	-2.4	1	-7.3	0	-7.3	0	1
4138	CAGAGCCG	486	-3.1	1	-3.1	1	-5.4	0	-5.4	0	1
4109	TGCCGGAC	602	-0.7	1	-3.1	1	-3.1	1	-5.7	0	1
4022	CCCCACAG	896	-0.3	1	-0.3	1	-8.9	0	-8.9	0	1
3857	GGCCACTG	931	0.5	1	-1.9	1	-5.1	0	-5.1	0	1
3675	GGTCCCAG	640	-3.4	1	-5.2	0	-5.2	0	-5.2	0	1
3526	GGATCGGC	245	1.0	1	0.8	1	-3.9	1	-3.9	1	1
3207	GGAGTGGG	399	-0.9	1	-0.9	1	-3.3	1	-3.3	1	1
3174	GGCTGCAC	82	-1.1	1	-1.1	1	-2.1	1	-2.1	1	1
3113	CCTGTGCC	717	-3.1	1	-3.1	1	-4.1	1	-4.1	1	1
3091	CCACCCAG	443	-3.1	1	-3.4	1	-5.4	0	-5.4	0	1
3067	GCAGCCAG	801	-0.8	1	-0.8	1	-4.3	1	-4.3	1	1
3059	GCCAGCAC	68	-0.8	1	-0.8	1	-3.4	1	-3.4	1	1
2965	GGCTGCAC	82	-4.5	1	-5.4	0	-5.4	0	-5.4	0	1
2928	GGCTGTGG	923	-2.1	1	-3.3	1	-5.4	0	-5.4	0	1
2779	CTGGTGGG	897	-1.1	1	-1.1	1	-6.0	0	-6.0	0	1
2630	GCCTCTGG	428	-0.6	1	-0.6	1	-2.6	1	-2.6	1	1
2575	GGGGCAAC	217	-3.8	1	-3.8	1	-3.8	1	-3.8	1	1
2462	GACTCGGG	851	-2.5	1	-2.5	1	-2.5	1	-3.7	1	1
2435	GCTTCCCC	181	-0.6	1	-0.6	1	-3.8	1	-3.8	1	1
2419	GTTGCCCC	702	-3.8	1	-3.8	1	-3.8	1	-3.8	1	1
2355	CCAGAGGC	913	-0.6	1	-0.6	1	-2.6	1	-2.6	1	1
2207	CCCACCAG	412	-1.1	1	-1.1	1	-5.6	0	-6.0	0	1
2057	CCACAGCC	438	-2.1	1	-3.3	1	-5.4	0	-5.4	0	1
1926	GTGCTGGC	553	-0.8	1	-0.8	1	-2.6	1	-3.4	1	1
1894	CTGGAGGG	891	-3.4	1	-3.4	1	-5.4	0	-5.4	0	1
1872	GGCACAGC	232	-3.1	1	-3.1	1	-4.1	1	-4.1	1	1
1811	GTGCAGCC	567	-1.1	1	-1.1	1	-2.1	1	-2.1	1	1
1778	CCCACTCC	884	-0.9	1	-0.9	1	-0.9	1	-3.3	1	1
1459	GCCGATCC	730	1.0	1	0.8	1	-0.6	1	-3.9	1	1
1310	CTGGGACC	155	-5.2	0	-5.2	0	-5.2	0	-5.2	0	1
1128	CAGTGGCC	446	0.5	1	-1.9	1	-5.1	0	-5.1	0	1
963	CTGTGGGG	411	-0.3	1	-0.3	1	-8.9	0	-8.9	0	1
876	GTCCGGCA	117	-3.1	1	-3.1	1	-5.1	0	-5.1	0	1
680	GAGGGCTC	270	-2.4	1	-2.4	1	-7.3	0	-7.3	0	1
660	CCAGGCAG	442	-1.1	1	-1.1	1	-2.4	1	-2.4	1	1
627	GGGCAAGG	176	-1.0	1	-3.6	1	-5.6	0	-5.6	0	1
519	GTGGCCAG	88	-3.6	1	-3.6	1	-6.0	0	-6.0	0	1
369	CAGGGAGC	463	-0.3	1	-0.9	1	-1.4	1	-1.4	1	1
4686	GCAGCAGC	235	-2.2	1	-2.2	1	-6.3	0	-6.3	0	0
4661	GTCCGGCA	117	-5.3	0	-6.3	0	-6.3	0	-6.3	0	0
4510	CCCCAGAC	580	-2.1	1	-5.1	0	-5.1	0	-5.1	0	0
4269	GCACTGCC	263	3.7	1	-7.3	0	-7.3	0	-7.3	0	0
3574	GCTCCAGG	834	-6.1	0	-6.1	0	-6.1	0	-6.1	0	0
2598	GGCCACTG	931	-5.9	0	-7.2	0	-7.2	0	-7.2	0	0
2550	GGGGAAGC	666	-0.6	1	-0.6	1	-3.8	1	-3.8	1	0
2523	CCCGAGTC	366	-2.5	1	-2.5	1	-3.6	1	-3.6	1	0
2387	CAGTGGCC	446	-5.9	0	-7.2	0	-7.2	0	-7.2	0	0
2020	GTGCAGCC	567	-4.5	1	-5.4	0	-5.4	0	-5.4	0	0
1612	CCTGGAGC	349	-6.1	0	-6.1	0	-6.1	0	-6.1	0	0
847	CGGCTCTG	1	-3.1	1	-3.1	1	-3.1	1	-5.4	0	0
716	GGCAGTGC	748	3.7	1	-7.3	0	-7.3	0	-7.3	0	0
475	GTCTGGGG	95	-2.1	1	-5.1	0	-6.0	0	-6.0	0	0
356	TGCCGCAC	85	-4.8	1	-4.8	1	-5.7	0	-5.7	0	0
324	TGCCGGAC	602	-5.3	0	-6.3	0	-6.3	0	-6.3	0	0
299	GCTGCTCC	720	-2.2	1	-2.2	1	-6.3	0	-6.3	0	0
289	CTCTCGCC	338	-0.9	1	-2.2	1	-2.5	1	-10.3	0	0

Primer location, sequence, name, eOST predicted free energy (ΔG) value using the indicated template window size (kcal/mol), eOST prediction of reaction success (1) or failure (0), and OST experimental results are indicated.

FASTA format is accessible via the web (URL http://www.hgsc.bcm.tmc.edu/~gmei/susan). Subsequently, eOST automatically calculates the stability of predicted secondary structures within the template in the vicinity of the octamer-binding site. The time consumed for each trace file varies with both the length of the sequence file and the number of candidate primers present within the template. On average, it takes between 30 and 60 s for each template to finish primer selection. The output files contain the primer library identification number, the primer sequence, the position of the 5′ base in the primer within the template, the strand with which the primer will pair, the number of times this primer occurs in the template, the PHRED values for the template-pairing bases, and the ΔG value of the template segment(s). The web-accessible version of eOST provides this information with the exception of the PHRED-related analysis.

Determining the threshold ΔG value

To increase OST sequencing success rate, eOST was also used to identify and optimize several parameters, including the base quality threshold, the optimal window size of the template sequence, and the threshold ΔG value, as criteria to identify octamers that are most likely to bind to the template and generate high quality sequencing data. We have refined parameters

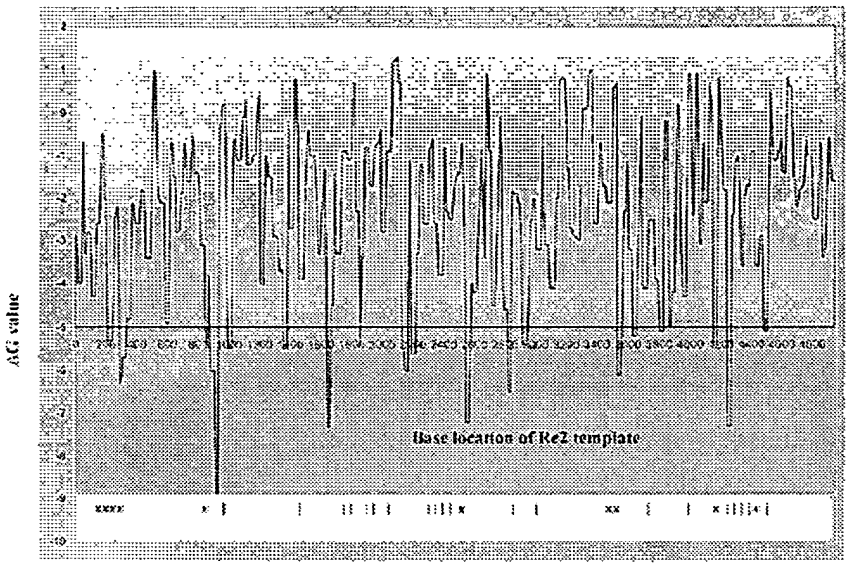

Figure 2. Graphic of ΔG values determined across the top strand of the Re2 sequence. The value for the first 50 base window was determined, the window was moved 10 bases, and the value for the next 50 base window was calculated. This process was continued until the entire sequence was analyzed. The ΔG value (kcal/mol) is plotted on the *y*-axis. | indicates the position of a successful primer and *x* indicates the position of a failed primer.

to improve the correlation between the stability of the secondary structure of the template around the octamer-binding site and the success or failure rate of the sequencing reaction. We demonstrate that $\Delta G > -5.0$ kcal/mol is a good predictor of OST reaction success. This user-set value may be changed as additional data are collected.

Currently, the simple dynamic folding algorithm in the eOST software is used to calculate the free energy and to predict the secondary structure of a DNA segment. The data indicate that the free energy value is useful for predicting the impact of a potential DNA template secondary structure on primer extension ability. First, a correlation between the calculated ΔG value and the success rate of primer extension was observed; Re2 analysis suggests a threshold value of −5 kcal/mol for predicting OST reaction outcome. Secondly, the site-specific binding efficiency of primers with double binding sites located in the same template implies that the secondary structure of the template plays an important role in either (or both) binding of primer to the specific site or extension by the DNA polymerase. Thirdly, by applying a simple folding algorithm to calculate the free energy across the Re2 sequence, some regions that have a low free energy and high potential to form secondary structure are located and associated with increased reaction failure (Fig. 2). However, it must be noted that some octamer-binding sites either directly associated with strong predicted secondary structures or located in template regions predicted to form secondary structures are able to prime sequence data. Experiments are in progress to understand this observation.

Identifying optimal template window size for ΔG calculation

As mentioned, a segment of DNA sequence can theoretically be folded in many different ways and each folding stage is associated with a defined free energy value. The lower the energy value, the greater the stability of the secondary structure for that segment. The eOST program lists the lowest free energy value for a particular template segment containing each primer-binding site in the output file. Each template segment of 50 bases encompassing the primer-binding site is selected and the free energy values for the different folding stages are calculated. The current version of the eOST program calculates these values by determining both nearest neighbor (11) and loop contributions (as per OLIGO 5.0; Molecular Biology Insights Inc.). In fact, by decreasing the template segment

Table 3. Determination of the overall predictive value using different template window sizes

	Reaction Prediction	No. of Successful OST Reactions	No. of Failed OST Reactions	Percent Correct Prediction
Window size 35				
> -5 kcal/mol (Predicted success)	56	44	12	79%
< -5 kcal/mol (Predicted failure)	7	1	6	86%
Window size 50				
> -5 kcal/mol (Predicted success)	49	42	7	86%
< -5 kcal/mol (Predicted failure)	14	3	11	79%
Window size 75				
> -5 kcal/mol (Predicted success)	28	24	4	86%
< -5 kcal/mol (Predicted failure)	35	21	14	40%
Window size 100				
> -5 kcal/mol (Predicted success)	24	22	2	92%
< -5 kcal/mol (Predicted failure)	39	23	16	41%

Analysis summary of the Re2 sequencing project using different segment window sizes. Raw data with ΔG values are shown in Table 2. Reaction prediction indicates the number of reactions (of the 63 Re2 reactions) that are predicted to succeed or fail, based on the indicated template window size. No. of successful OST reactions indicates the number of reactions (of 63) in which the octamer primed the reaction. No. of failed OST reactions indicates the number of reactions (of 63) in which the octamer did not prime the reaction. Percent correct prediction is determined by dividing the number of successful OST reactions by the predicted number of successful reactions (e.g. for the 50 base window, 42/49 = 86%) or by dividing the number of failed OST reactions by the predicted number of failed reactions (11/14 = 79%).

Table 4. Re2 project summary of ΔG values determined using different template window placements

Primer Location	Primer Sequence	Primer Name	ΔG1 (-20)	ΔG2 (-10)	ΔG3 (0)	ΔG4 (+10)	ΔG5 (+20)	OST Experimental Result
1459	GCCGATCC	730	-2.7	0.8	0.8	-0.6	-0.8	1
3526	GGATCGGC	245	-0.8	-0.6	0.8	0.8	-2.7	1
4022	CCCCACAG	896	0.2	-0.3	-0.3	-8.9	-8.9	1
963	CTGTGGGG	411	-8.9	-8.9	-0.3	-0.3	0.2	1
2355	CCAGAGGC	913	-0.8	-0.6	-0.6	-2.6	-2.6	1
2630	GCCTCTGG	428	-2.6	-2.6	-0.6	-0.6	-0.8	1
2435	GCTTCCCC	181	-3.8	-3.8	-0.6	-2.3	-2.5	1
3067	GCAGCCAG	801	-1.7	-0.8	-0.8	-0.7	-3.4	1
1926	GTGCTGGC	553	-3.4	-0.8	-0.8	-0.8	-1.7	1
3059	GCCAGCAC	68	-1.7	-0.8	-0.8	-0.8	-3.4	1
4616	GCTCCCTG	948	-0.9	-0.9	-0.9	-0.6	-1.4	1
1778	CCCACTCC	884	-0.5	-0.9	-0.9	-0.9	-1.1	1
369	CAGGGAGC	463	-1.4	-0.6	-0.9	-0.9	-0.9	1
3207	GGAGTGGG	399	-1.1	-0.9	-0.9	-0.9	-0.5	1
4325	CTGCCTGG	927	-2.4	-1.1	-1.1	0.6	-1.0	1
2778	CTGGTGGG	897	-5.6	-5.6	-1.1	-1.7	-6.0	1
660	CCAGGCAG	442	-1.0	0.6	-1.1	-1.1	-2.4	1
2207	CCCACCAG	412	-6.0	-1.7	-1.1	-5.6	-5.6	1
3857	GGCCACTG	931	-5.1	-5.1	-1.9	-0.2	-0.2	1
1128	CAGTGGCC	446	-0.2	-0.2	-1.9	-5.1	-5.1	1
4696	GGCGAGAG	823	-6.3	-2.2	-2.2	-2.5	-4.3	1
4305	GAGCCCTC	755	-7.3	-7.3	-2.4	-1.1	-1.1	1
680	GAGGGCTC	270	-1.1	-1.1	-2.4	-7.3	-7.3	1
2462	GACTCGGG	851	-2.3	-2.5	-2.5	-2.5	-2.5	1
3113	GCTGTGCC	717	-3.1	-3.1	-3.1	-4.1	-4.1	1
4109	TGCCGGAC	602	-3.1	-3.1	-3.1	-2.1	-2.1	1
4138	CAGAGCCG	486	-3.8	-3.1	-3.1	-3.1	-1.7	1
1872	GGCACAGC	232	-4.1	-4.1	-3.1	-3.1	-3.1	1
876	GTCCGGCA	117	-2.1	-2.1	-3.1	-3.1	-3.1	1
2928	GGCTGTGG	923	-2.1	-2.1	-3.3	-3.3	-5.4	1
2057	CCACAGCC	438	-5.4	-3.3	-3.3	-2.1	-2.1	1
1894	CTGGAGGG	891	-2.3	-4.1	-3.4	-3.4	-0.8	1
3091	CCCTCCAG	406	-0.8	-3.4	-3.4	-4.1	-2.3	1
4358	CCTTGCCC	661	-1.0	-1.0	-3.6	-3.6	-3.6	1
4466	CTGGCCAC	573	-3.6	-3.6	-3.6	-2.9	-2.9	1
627	GGGCAAGG	176	-3.6	-3.6	-3.6	-1.0	-1.0	1
519	GTGGCCAG	88	-2.9	-2.9	-3.6	-3.6	-3.6	1
2410	GTTGCCCC	702	-3.8	-3.8	-3.8	-3.8	-0.8	1
2575	GGGGCAAC	217	-0.8	-3.8	-3.8	-3.8	-3.8	1
4619	GTGCGGCA	570	-2.1	-4.8	-4.8	-4.8	-5.7	1
3675	GGTCCCAG	640	-5.2	-5.2	-5.2	-3.4	-3.4	1
1310	CTGGGACC	155	-3.4	-3.4	-5.2	-5.2	-5.2	1
2550	GGGGAAGC	666	-2.5	-2.3	-0.6	-3.8	-3.8	0
299	GCTGCTCC	720	-2.5	-2.2	-2.2	-6.3	-6.3	0
289	CTCTCGCC	338	-4.3	-2.5	-2.2	-2.2	-6.3	0
4686	GGAGCAGC	235	-6.3	-6.3	-2.2	-2.2	-2.5	0
2523	CCCGAGTC	366	-2.5	-2.5	-2.5	-2.5	-2.3	0
847	CGGCTCTG	1	-1.7	-3.1	-3.1	-3.1	-3.8	0
366	TGCCGCAC	85	-5.7	-4.8	-4.8	-4.8	-2.1	0
4510	CCCCAGAC	580	-2.9	-2.9	-5.1	-5.1	-5.1	0
475	GTCTGGGG	95	-5.1	-5.1	-5.1	-2.9	-2.9	0
2020	GTGCAGCC	567	-4.5	-4.5	-5.4	-5.4	-5.4	0
2965	GGCTGCAC	82	-5.4	-5.4	-5.4	-4.5	-4.5	0
324	TGCCGGAC	602	-3.2	-6.3	-6.3	-5.7	-5.7	0
4661	GTCCGGCA	117	-5.7	-5.7	-6.3	-6.3	-3.2	0
2598	GGCCACTG	931	-7.2	-7.2	-7.2	-5.9	-4.2	0
2387	CAGTGGCC	446	-4.2	-5.9	-7.2	-7.2	-7.2	0
716	GGCAGTGC	748	-7.3	-7.3	-7.3	-1.8	-1.8	0
4269	GCACTGCC	263	-1.8	-1.8	-7.3	-7.3	-7.3	0

Primer location, sequence, and name are shown. ΔG values were calculated for 50 base windows of the template centered on the octamer ($\Delta G3$) or shifted as described in Figure 3. The experimental results are shown as 1 and 0, where 1 indicates reaction success and 0 indicates reaction failure. Data are sorted by experimental results, and then by increasing strength of the predicted secondary structures ($\Delta G3$ value).

window size from 100 to 50 bases, eOST increases the predictive value for the Re2 project (Table 3). Specifically, the percentage of reactions that were predicted to fail but were successful decreased from 59 to 21%. Although the predicted success rate decreased slightly (from 92 to 86%), the number of reactions predicted to succeed increased from 24 to 49.

The template window can be moved relative to the octamer-binding site. In fact, we have observed a significant change in

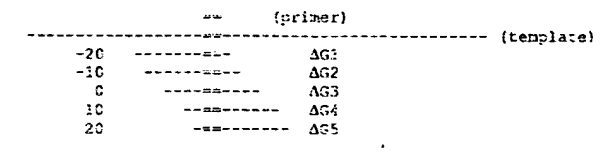

Figure 3. Schematic view of window placement. The ΔG values shown in Table 4 were calculated using 50 base template windows centered on the octamer ($\Delta G3$), that shifted the window 20 or 10 bases left of this site ($\Delta G1$ or $\Delta G2$, respectively) or that shifted the window 10 or 20 bases right of this site ($\Delta G4$ or $\Delta G5$, respectively).

the free energy value by shifting the 50 base window 10 or 20 bases to the right or left (Fig. 3 and Table 4). These free energy value changes should allow us to determine the best way to position the window for analysis. Future work can determine whether a better predictor of reaction outcome is obtained by averaging values determined for particular window positions.

Uses for eOST

In our previous report, an octamer library provided sufficient coverage to sequence both genomic and cDNA sequences (8). This may now be especially important for gap closure in genomic templates, given the shift in Human Genomic sequencing strategy brought about by private sequencing efforts. Effectively, this shift promotes increased throughput of new sequence information with a corresponding decrease in sequence confidence. Therefore, investigators interested in a particular genomic region may need to confirm the sequence before proceeding with their experiments. Additionally, as genome projects near completion, eOST could be used as a rapid and cost-effective alternative for gap closure using octamer-appropriate templates, and in the finishing phase of genomic DNA sequencing. eOST can be modified to analyze multiple sequences in parallel. eOST and OST, when combined, will allow researchers to accomplish these needs rapidly and relatively inexpensively.

The characterization of gene information in the Human Genome Project relies heavily on expressed sequence tags (ESTs). ESTs are generated from a single DNA sequencing pass from either or both ends of a cloned cDNA. They are typically short (average 200–400 bp), generally of lower-quality data, and only contribute limited information about each clone. As the pace of hunting for and analyzing human genes accelerates, large-scale efforts to systematically sequence human full-length cDNA clones are reported (14,15). eOST could also be incorporated into large-scale sequencing of cDNA clones in the final assembly stage. Additionally, eOST could be used to manage a library comprised of longer primers or to determine whether our observation of the impact of template secondary structures applies to longer primers.

ACKNOWLEDGEMENTS

The authors are grateful to Paul Hardin, Wei Yu, and Anelia Kraltcheva for critically reading the manuscript, and to Hardin laboratory members for helpful discussions. This work was supported by the NIH (grant R29-HG01151 to S.H.H.).

REFERENCES

1. Kaiser,R.J., MacKellar,S.L., Vinayak,R.S., Sanders,J.Z., Saavedra,R.A. and Hood,L.E. (1989) *Nucleic Acids Res.*, **17**, 6087–6102.
2. Li,P., Kupfer,K.C., Davis,C.J., Burbee,D., Evans,G.A. and Garner,H.R. (1997) *Genomics*, **40**, 476–485.
3. Studier,F.W. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 6917–6921.
4. Siemieniak,D.R. and Slightom,J.L. (1990) *Gene*, **96**, 121–124.
5. Burbelo,P.D. and Iadarola,M.J. (1994) *Biotechniques*, **16**, 645–650.
6. Hardin,S.H., Jones,L.B., Homayouni,R. and McCollum,J.C. (1996) *Genome Res.*, **6**, 545–550.
7. Jones,L.B. and Hardin,S.H. (1998) *Nucleic Acids Res.*, **26**, 2824–2826.
8. Jones,L.B. and Hardin,S.H. (1998) *ABRF News*, **9**, 6–10.
9. Ewing,B. and Green,P. (1998) *Genome Res.*, **8**, 186–194.
10. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) *Genome Res.*, **8**, 175–185.
11. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
12. Rychlik,W. and Rhoads,R.E. (1989) *Nucleic Acids Res.*, **17**, 8543–8551.
13. Lvovsky,L., Ioshikhes,I., Raja,M.C., Zevin-Sonkin,D., Sobolev,I.A., Liberzon,A., Shwartzburd,J. and Ulanovsky,L.E. (1998) *Nucleic Acids Res.*, **26**, 5525–5532.
14. Ishikawa,K., Nagase,T., Nakajima,D., Seki,N., Ohira,M., Miyajima,N., Tanaka,A., Kotani,H., Nomura,N. and Ohara,O. (1997) *DNA Res.*, **31**, 307–313.
15. Yu,W., Andersson,B., Worley,K.C., Muzny,D.M., Ding,Y., Liu,W., Ricafrente,J.Y., Wentland,M.A., Lennon,G. and Gibbs,R.A. (1997) *Genome Res.*, **7**, 353–358.